A priori par normes mixtes pour les problèmes inverses

Application à la localisation de sources en M/EEG

Matthieu Kowalski*, Alexandre Gramfort**

* *L2S, UMR 8506 CNRS - SUPELEC - Univ Paris-Sud 11*
Plateau de Moulon
91192 Gif-sur-Yvette Cedex, France
matthieu.kowalski@lss.supelec.fr
+33 (0)1 69 85 17 47

** *INRIA, Projet PARIETAL*
NeuroSpin CEA Saclay
Bat. 145, PC 156
91192 Gif-sur-Yvette, France
alexandre.gramfort@inria.fr
+33 (0)1 69 08 80 82

RÉSUMÉ. *On s'intéresse aux problèmes inverses sous déterminés, et plus particulièrement à la localisation de sources en magnéto et électro-encéphalographie (M/EEG). Bien que l'on ait à disposition un modèle physique de la diffusion (ou du "mélange") des sources, le caractère très sous-déterminé de ces problèmes rend l'inversion difficile. La nécessité de trouver des a priori forts et pertinents physiquement sur les sources est une des difficultés. Dans le cadre du problème inverse de la M/EEG, la parcimonie classique mesurée par une norme ℓ_1 n'est pas suffisante, et donne des résultats non réalistes. On propose ici de prendre en compte une parcimonie structurée grâce à l'utilisation de normes mixtes – notamment d'une norme mixte sur trois niveaux. La méthode est utilisée sur des signaux MEG issus d'expériences de stimulation somesthésique. Lorsqu'ils sont stimulés, les différents doigts de la main activent des régions distinctes du cortex sensoriel primaire. L'utilisation d'une norme mixte à trois niveaux permet d'injecter cet a priori dans le problème inverse et ainsi de retrouver la bonne organisation corticale des zones actives. Nous montrons également que la méthode la plus classiquement utilisée dans le domaine échoue dans cette tâche.*

2e soumission à *Traitement du Signal*, le 25/01/2010

ABSTRACT. *We are interested by under-determined inverse problems, and more specifically by source localization in magneto and electro-encephalography (M/EEG). Although there is a physical model for the diffusion (or "mixing") of the sources, the (very) under-determined nature of the problem leads to a difficult inversion. The need for strong and physically relevant priors on the sources is one of the challenge. For M/EEG classical sparsity prior based on the l1 norm is not adapted, and gives unrealistic results. We propose to take into account a structured sparsity thanks to the use of mixed norms, especially a mixed norm with three indices. The method is then applied on MEG signals obtained during somesthetic stimulation. When stimulated, hand fingers activate separate regions of the primary somatosensory cortex. The use of a three level mixed norm allows to take this prior into account in the inverse problem in order to correctly recover the organization of associated brain regions. We also show that classical methods fail for this task.*

MOTS-CLÉS : *normes mixtes, problème inverse, opérateurs de proximité, électroencéphalographie, magnétoencéphalographie*

KEYWORDS: *mixed norms, inverse problems, proximal operators, electroencephalography, magnetoencephalography*

Abstract

We are interested by under-determined inverse problems, and more specifically by source localization in magneto and electro-encephalography (M/EEG). Although there is a physical model for the diffusion (or "mixing") of the sources, the (very) under-determined nature of the problem leads to a difficult inversion. Such a problem can be written as

$$M \approx GX \ ,$$

where M is the measurement matrix, G the lead-field matrix (a.k.a. gain matrix), and X the amplitudes of the sources that need to be estimated.

The problem can be cast into the classical approach of regularized least-squares. A solution of the inverse problem is obtained via the minimization of a convex cost function, constructed with a ℓ_2 data fidelity term to which is added a relevant regularization term, also denoted *prior*. Designing strong and physically relevant priors on the sources while providing efficient algorithms to use them on realistic datasets is the challenge addressed in this contribution.

The variational formulation can be expressed as

$$\hat{X} = \arg\min_X \frac{1}{2}\|M - GX\|^2 + \lambda P(X) \ , \lambda > 0 \ ,$$

where P is a convex prior on the sources. Classical choices in the context of M/EEG are the ℓ_2 minimum energy prior and the ℓ_1 sparse prior. However, such priors suffer from several limitations. To go beyond such ℓ_p norms, the use of a structured prior is particularly adapted. To do so, we propose to use priors based on mixed norms, and, more precisely, we detail the ℓ_{212} three level mixed norm defined as

$$\|X\|_{212} = \left(\sum_i \left(\sum_k \left(\sum_t |X_{ikt}|^2 \right)^{1/2} \right)^2 \right)^{1/2} \ .$$

Such a prior is based on the concepts of "diversity" and "sparsity" that can be used to obtain structured solutions. The motivation for introducing such a three level mixed norm in M/EEG, is to integrate a priori knowledge between multiple "experimental conditions". Indeed, if index i represents the source location, k the experimental condition and t the time, the ℓ_{212} prior promotes source estimates where a given source is active only for few conditions. The consequence is that the overlap of activations between the different conditions is reduced.

Another contribution of this work is to present efficient algorithms, primarily developed in the convex optimization community, to optimize the cost functions under consideration. More precisely, the algorithms presented rely on the computation of the so-called proximity operators associated with the different priors under consideration. Iterative thresholding algorithms such as FISTA are general algorithms which can be used to minimize cost functions formed by a least-square term and a convex prior

term P, that need not be necessarily differentiable. This is essential in order to use mixed-norm based a non differentiable sparse prior such ℓ_1, ℓ_{21} or ℓ_{212}. Moreover, we discuss the possible criteria used to stop such iterative algorithms. We propose to use a stopping criterion based on duality gaps. Derivation of this criterion for the mixed norms considered is presented.

The method is finally applied on MEG signals obtained during somesthetic stimulation. When stimulated, hand fingers activate separate regions of the primary somatosensory cortex. The use of the ℓ_{212} three level mixed norm allows to take this prior into account in the inverse problem in order to correctly recover the organization of associated brain regions. We also show that classical methods fail for this task.

1. Introduction

De nombreux problèmes de traitement du signal, comme la séparation de sources, la tomographie, ou – problème considéré plus particulièrement ici – la localisation de sources en M/EEG, consistent à restituer un ou plusieurs signaux à partir d'observations incomplètes et bruitées. De tels problèmes sont qualifiés de problèmes inverses « mal posés ».

Les exemples précédents sont en effet des problèmes sous déterminés, car on cherche à reconstruire plus de signaux qu'on en observe. La modélisation d'un tel problème s'effectue généralement en deux étapes :

1) Trouver un bon modèle pour le problème direct, qui permet de lier les signaux d'entrées aux sorties mesurées ;

2) Inverser le modèle précédent, c'est à dire estimer les signaux d'entrées à partir des seules mesures à disposition.

On s'intéresse ici seulement à la deuxième étape, c'est à dire à l'inversion proprement dite. On supposera qu'on a à disposition un modèle linéaire permettant de lier les signaux d'entrées aux mesures effectuées, qui pourra s'écrire sous la forme

$$M \simeq GX \ ,$$

où M représente les signaux mesurés et G est un opérateur linéaire permettant de relier les signaux d'intérêt X aux mesures. L'opérateur G dans le contexte M/EEG est communément appelé *matrice de gain*. Le problème considéré étant sous déterminé, la matrice G a moins de lignes que de colonnes. L'inversion d'un tel problème n'est alors possible que si l'on dispose d'*a priori* sur la solution recherchée.

Les principales contributions de ce travail sont les suivantes. On généralise les normes mixtes à trois niveaux et plus, en expliquant l'intérêt de telles normes. On illustre l'utilisation d'une norme mixte sur trois niveaux sur un problème de localisation de sources dans le cadre des signaux M/EEG. Les résultats obtenus, aussi bien sur des données synthétiques que sur un problème réel, montrent que l'introduction d'a priori structurés permet d'améliorer les performances par rapport à l'état de l'art. L'état de l'art en M/EEG est principalement basé sur des normes à un seul niveau (Dale *et al.*, 1993; Baillet *et al.*, 2001; Gramfort, 2009). Toutefois, depuis plus récemment, il a été proposé une norme à 2 niveaux (Ou *et al.*, 2009) pour laquelle nous fournissons une méthode d'optimisation nettement plus efficace que celle alors proposée.

La structure du reste de l'article est la suivante. La section 2 rappelle le problème direct et présente l'estimation de solutions comme la résolution d'un problème d'optimisation. On rappelle les principaux *a priori* utilisés, comme celui de moindre énergie. La section 3 rappelle la définition des normes mixtes sur deux niveaux et les généralise à trois niveaux, puis à un nombre de niveaux quelconque. La résolution du problème d'optimisation convexe considéré par des méthodes de premier ordre efficaces – dont on peut démontrer l'optimalité, en un certain sens, en terme de vitesse de convergence

– est exposée dans la section 4. Enfin, la section 5 illustre notre approche sur un problème de localisation de sources dans le cadre de signaux M/EEG.

2. Approche générale et état de l'art

Le but d'un problème inverse en traitement du signal est de restituer un ou plusieurs signaux à partir d'observations souvent incomplètes et bruitées. Les signaux sont supposés liés par un modèle physique. On supposera ici que l'on dispose de N capteurs pour mesurer des signaux de longueur T. Les mesures, que l'on représentera par une matrice $M \in \mathbb{R}^{N \times T}$, sont supposées issues d'un mélange de I sources de longueur T représentées par une matrice $X \in \mathbb{R}^{I \times T}$. Le mélange est effectué par un opérateur linéaire $G \in \mathbb{R}^{N \times I}$. Cet opérateur G représente une transformation physique liant les sources X et les observations M :

$$M = GX + b\,, \qquad [1]$$

où b représente un bruit additif.

On se place dans le cadre où, bien que l'on dispose d'un modèle réaliste pour l'opérateur G, le nombre N d'observations à notre disposition est très inférieur au nombre I de sources à estimer. En particulier, on s'intéresse au problème de localisation de sources en M/EEG, pour lequel $N \simeq 100$ et I peut aller de 5000 à 30000, selon la précision du modèle utilisé (Dale *et al.*, 1993). Le but de ce problème inverse est d'estimer la position et l'amplitude des sources de courant dans le cerveau à partir d'une centaine de mesures des champs électrique ou magnétique induits par l'activité neuronale jusqu'à l'extérieur de la tête. Si ce problème est à rapprocher de la séparation de sources audio, son caractère très sous-déterminé ne permet pas d'utiliser les techniques classiques habituellement mise en oeuvre (voir par exemple (Vincent *et al.*, 2007) et les références pour un aperçu des méthodes utilisées en séparation de sources audio).

L'exemple de la localisation de sources en M/EEG nous servira de fil conducteur tout au long de cet article. Il nous permettra d'illustrer le propos afin de rendre plus intuitif les concepts utilisés. Ces derniers restent toutefois très généraux et peuvent être utilisés pour la résolution d'un problème inverse se formulant comme [1].

2.1. *Approche générale : une formulation variationnelle*

Une approche classique est de rechercher une estimation des sources X comme l'optimum d'une fonctionnelle. Les fonctionnelles généralement utilisées peuvent s'écrire sous la forme :

$$\mathcal{F}(X) = f_1(X) + \lambda f_2(X), \qquad [2]$$

et font apparaître :

– Un terme d'attache aux données, f_1, qui quantifie à quel point les signaux à estimer expliquent les mesures. Ce terme permet de prendre en compte le bruit de mesure.

– Un terme de régularisation, f_2, qui permet d'introduire un *a priori* sur les sources recherchées. Cet *a priori* est indispensable dans le cas d'un problème inverse sous-déterminé, en raison de l'existence d'une infinité de solutions pouvant expliquer les mêmes mesures.

– Un paramètre $\lambda > 0$, qui permet de régler l'importance du terme d'attache aux données f_1 par rapport au terme de régularisation f_2. Intuitivement, plus λ est choisi « petit », plus le terme d'attache aux données sera prépondérant et plus les signaux expliqueront bien les mesures. Inversement, si λ est choisi « grand » la solution sera fortement régularisée.

On se restreint ici aux cas où f_1 et f_2 sont des fonctions convexes. En effet, les travaux effectués par la communauté de l'optimisation convexe sont abondants, et il existe aujourd'hui des algorithmes efficaces pour optimiser de telles fonctionnelles. Les détails liés à l'optimisation proprement dite sont présentés dans la section 4.

Le terme d'attache aux données choisi ici est une norme ℓ_2 :

$$f_1(X) = \frac{1}{2}\|M - GX\|_2^2. \qquad [3]$$

Ce terme a pour but de minimiser l'énergie du résidu existant entre l'estimation d'une solution et les mesures. D'un point de vue Bayesien, minimiser [3], revient à maximiser une vraisemblance du type $e^{-\frac{1}{2}\|M-GX\|_2^2}$, ce qui est équivalent à un *a priori* gaussien sur le bruit b. Notons qu'on appelle ici norme ℓ_2 ce qui correspond en fait à la norme de Frobenius d'une matrice. Remarquons aussi que le facteur $\frac{1}{2}$ n'est ici que pour des raisons calculatoires, afin d'annuler le facteur 2 issu de la dérivée du carré.

Il existe d'autres types d'attache aux données, comme la minimisation de la norme infinie du résidu. Cette dernière a pour avantage de borner uniquement la plus grande valeur (en valeur absolue) du résidu. Les autres coefficients sont laissés « libres ». Une telle attache aux données a été étudiée par (Candès *et al.*, 2005). La principale limitation vient de la difficulté à optimiser les fonctionnelles obtenues. On se restreint donc dans ce travail à un terme d'attache aux données ℓ_2, ce qui est totalement pertinent dans le cadre de la M/EEG. Ce problème inverse est donc classiquement présenté sous une formulation variationnelle du type :

$$X^* = \arg\min_X \frac{1}{2}\|M - GX\|_2^2 + \lambda f_2(X)\ , \lambda \in \mathbb{R}_+\ . \qquad [4]$$

On discute dans le reste de cette section différents choix possibles pour f_2. Remarquons que la formulation précédente peut s'écrire de façon équivalente sous la forme (Weiss, 2008) :

$$X^* = \arg\min_X f_2(X) \qquad [5]$$

$$\text{sous contrainte} \quad \|M - GX\|_2 \leq \delta,\ \delta \in \mathbb{R}_+\ . \qquad [6]$$

Cette dernière formulation semble plus naturelle, car si le bruit est effectivement un bruit blanc gaussien, on peut estimer relativement aisément le paramètre δ. Cependant, même s'il existe des algorithmes permettant d'optimiser [4] (Weiss, 2008), ces

derniers sont en pratique souvent moins rapides, notamment sur l'application considérée ici. Par conséquent, on se restreint dans ce travail à la formulation [4], qui est actuellement la plus répandue.

2.2. *Régularisation par normes* $\ell_{\mathbf{w};p}$

Un choix classique pour le terme de régularisation f_2 est une norme $\ell_{\mathbf{w};p}$ (élevée à la puissance p) dont on rappelle la définition.

Définition (Norme $\ell_{\mathbf{w};p}$)**.** *Soit* $\mathbf{x} = (x_1, \ldots, x_n)$ *un vecteur de* $\mathbb{R}^n$ *et* $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n_+$, *un vecteur de poids strictement positifs. Soit* $p \geq 1$. *On appelle norme* $\ell_{\mathbf{w};p}$ *de* $\mathbf{x}$ *la quantité :*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} w_i |x_i|^p \right)^{1/p} .$$

Bien que cette définition soit donnée pour un vecteur, l'extension au cas matriciel est directe. Les deux normes les plus utilisées en pratique sont la norme $\ell_{\mathbf{w};2}$ et la norme $\ell_{\mathbf{w};1}$.

2.2.1. *La norme* $\ell_{\mathbf{w};2}$

La norme $\ell_{\mathbf{w};2}$, comme cela a déjà été mentionné pour le terme d'attache aux données, permet de mesurer l'énergie d'un vecteur. L'utilisation d'un tel terme régularisant a donc pour effet de chercher une solution d'énergie minimale :

$$\hat{X} = \arg\min_X \frac{1}{2}\|M - GX\|_2^2 + \frac{\lambda}{2}\|X\|_{\mathbf{w};2}^2 . \qquad [7]$$

D'un point de vu Bayesien, cela correspond à un *a priori* Gaussien sur la solution recherchée. Une conséquence directe d'un tel *a priori* est que l'énergie est « étalée » sur tous les coefficients. Ainsi, pour le problème considéré, un tel *a priori* conduit à l'estimation de régions cérébrales actives dont l'extension spatiale est souvent surestimée. La solution avec *a priori* ℓ_2 est connue dans la littérature M/EEG comme la solution "Minimum Norm" (Baillet *et al.*, 2001).

Afin de réduire ce biais d'estimation, une approche consiste à introduire de la parcimonie dans l'*a priori*, via notamment l'utilisation d'une norme $\ell_{\mathbf{w};1}$.

2.2.2. *La norme* $\ell_{\mathbf{w};1}$

Cette norme est connue pour favoriser des solutions parcimonieuses dans un problème de régression tel que celui proposé, et a connu un succès croissant durant les quinze dernières années, notamment avec les travaux sur l'échantillonnage compressif (Donoho, 2006). La parcimonie est une hypothèse forte, qui suppose que seul un nombre limité de coefficients intervient effectivement dans la solution recherchée.

Cet *a priori* est à utiliser avec précaution. Si l'hypothèse de parcimonie permet d'obtenir d'excellents résultats dans diverses applications comme le débruitage (Kowalski *et al.*, 2008; Févotte *et al.*, 2008; Dupé *et al.*, 2009), certains problèmes de séparation de sources (Bobin *et al.*, 2008; Kowalski *et al.*, 2009b) ou de codage (Daudet *et al.*, 2004), elle peut conduire à des solutions non réalistes si elle n'est pas satisfaite.

Dans l'application visée ici, son usage est inadapté. Une première raison est que l'hypothèse de parcimonie n'est pas toujours valide dans le contexte de la M/EEG. Certaines données en M/EEG sont produites par des regions cérébrales avec une certaine extension spatiale. Une deuxième raison, plus fondamentale, est que la norme ℓ_1 ne conduit généralement pas à une estimation de sources dont les décours temporels sont réguliers. Or, une source active au temps t a de fortes chances d'être active au temps $t+1$. Pour comprendre l'origine de cette observation, on peut remarquer que la régularisation ℓ_1 correspond d'un point de vu Bayesien à un *a priori* Laplacien sur la solution recherchée. En effet, minimiser une fonction du type $\lambda\|x\|_1$ revient à maximiser $\prod_k e^{-\lambda|x_k|}$, ce qui fait apparaître clairement l'hypothèse d'indépendance sous-jacente à la norme ℓ_1. Les coefficients de X sont estimés indépendamment dans l'espace et dans le temps. C'est cette indépendance qui empêche d'obtenir une solution regulière en temps, et donc réaliste. Une solution consiste à contraindre les coefficients associés à une même source à être estimés conjointement par un regroupement en temps. Ce type de contrainte, exposée dans la section 3, passe par l'utilisation d'une norme mixte à deux niveaux (Ou *et al.*, 2009).

2.3. *Autres approches*

Il existe bien d'autres types de régularisation que les normes $\ell_{\mathbf{w};p}$ et les normes mixtes exposées dans la section suivante. On pourra par exemple citer la variation totale, très utilisée en image. Il existe aussi d'autres approches que l'approche variationnelle avec optimisation d'une fonction convexe, notamment les approches Bayesiennes comme celles décrites dans (Wipf *et al.*, 2009). Des approches plus ad-hoc ont aussi été développées pour l'estimation de sources en M/EEG, comme LORETA (Pascual-Marqui *et al.*, 1994), dSPM (Dale *et al.*, 2000) ou MUSIC (Mosher *et al.*, 1992). Pour un état de l'art plus complet de la localisation de sources en M/EEG, on pourra se reporter aux chapitres 3 et 4 de la thèse (Gramfort, 2009).

Afin de palier l'indépendance supposée des coefficients associés aux normes $\ell_{\mathbf{w};p}$, on présente dans la section suivante une manière de structurer les coefficients grâce aux normes mixtes. Cette approche rentre dans le cadre général [2], et nous permettra d'utiliser les algorithmes d'optimisation convexe présentés dans la section 4.

3. Structures et parcimonie par les normes mixtes

On présente dans cette section les normes mixtes, qui seront utilisées comme fonction de régularisation (le terme f_2 dans [4]). On rappelle tout d'abord la définition des normes mixtes sur deux niveaux. On étend ensuite la définition sur trois niveaux, qui est le cadre de la norme utilisée dans les applications présentées dans la section 5. On généralise ensuite la définition d'une norme mixte sur n niveaux. Enfin, on donne la norme mixte duale ainsi que le conjugué au sens Fenchel.

3.1. *Normes mixtes sur deux niveaux*

Nous utiliserons les notations suivantes. Soit une suite $(\mathbf{x}) \in \mathbb{C}^{\mathbb{N}}$ indexée par un double indice $(g,m) \in \mathbb{N}^2$ telle que $(\mathbf{x}) = (x_{g,m})_{(g,m)\in\mathbb{N}^2}$. On peut alors considérer les deux sous suites canoniques $(\mathbf{x}_g) = (x_{g,1}, x_{g,2}, \ldots)$ pour g fixé, et $(\mathbf{x}_{.,m}) = (x_{1,m}, x_{2,m}, \ldots)$ pour m fixé.

La double indexation est purement conventionnelle et sert à introduire une dépendance entre les coefficients. Cette dépendance est utilisée dans la définition suivante d'une norme mixte.

Définition (Normes mixtes sur deux niveaux)**.** *Soit $(\mathbf{w}) \in \mathbb{R}^{\mathbb{N}}$, indicée par un double indice $(g,m) \in \mathbb{N}^2$, et telle que pour tout (g,m) $w_{g,m} > 0$.*

Soit $p \geq 1$ et $q \geq 1$. On appelle norme mixte (pondérée) de $(\mathbf{x}) \in \mathbb{C}^{\mathbb{N}}$, la norme $\ell_{\mathbf{w};p,q}$ définie par

$$\|\mathbf{x}\|_{\mathbf{w};p,q} = \left(\sum_g \left(\sum_m w_{g,m} |x_{g,m}|^p \right)^{q/p} \right)^{1/q}.$$

Les cas $p = +\infty$ et $q = +\infty$ sont obtenus en remplaçant la norme correspondante par le supremum.

La norme mixte $\ell_{\mathbf{w};p,q}$ peut être vue comme une composition des normes $\ell_{\mathbf{w};p}$ et ℓ_q :

$$\begin{aligned} \|\mathbf{x}\|_{\mathbf{w};p,q} &= \left(\sum_g \|\mathbf{x}_g\|_{\mathbf{w};p}^q \right)^{1/q} \\ &= \left\| \left(\sum_m W_{.,m} |\mathbf{x}_{.,m}|^p \right)^{1/p} \right\|_q , \end{aligned} \qquad [8]$$

où $|\mathbf{x}_{.,m}|^p$ est obtenu en prenant le module de chaque élément de la suite $(\mathbf{x}_{.,m})$ puis en les élevant à la puissance p ; et où $W_{.,m} = \text{diag}(w_{1,m}, \ldots, w_{g,m}, \ldots)$. Ce type

de notation sera utilisée par la suite s'il n'y a pas d'ambiguïté. On peut aussi remarquer facilement que les normes mixtes généralisent les normes ℓ_p usuelles si $p = q$: $\|\mathbf{x}\|_{\mathbf{w};p,p} = \|\mathbf{x}\|_{\mathbf{w};p}$. Lorsque $p < 1$ ou $q < 1$ on ne peut plus parler de normes au sens strict, mais on peut généraliser la définition pour obtenir des quasi-normes (en particulier, on perd l'inégalité triangulaire).

Les deux indices g et m peuvent être interprétés comme une hiérarchie sur les coefficients. La double indexation nécessaire à la définition de la norme mixte permet en effet de considérer les coefficients « par groupes ». Les coefficients sont séparés entre groupes « aveugles » les uns des autres, et les coefficients d'un même groupe interagissent entre eux (et sont corrélés d'une certaine manière). Avec certaines normes mixtes, les coefficients d'un même groupe « s'uniront » tandis qu'avec un autre choix, ces coefficients entreront en compétition. Avec ces notations, l'indice g représente l'indice du groupe, et l'indice m l'indice de membre d'un groupe. Les normes mixtes sont donc un moyen d'introduire explicitement un couplage entre les coefficients à la place de l'hypothèse d'indépendance qui est derrière les normes ℓ_p. Ce couplage est dépendant des choix faits pour p et q.

Selon la valeur de p, les normes ℓ_p mesurent la diversité ou la parcimonie : pour des petites valeurs de p, la norme ℓ_p mesure la diversité, tandis qu'elle mesure la parcimonie pour des grandes valeurs. On rappelle qu'un vecteur est parcimonieux si son nombre de coefficients non-nuls est faible. Les normes mixtes permettent ainsi de mélanger ces deux concepts en jouant sur p et q. Dans un contexte de régression, les normes mixtes favorisent des types spécifiques de parcimonie et de diversité jointes.

Les normes mixtes étant évidemment des normes, la proposition suivante précise les cas de convexité et de convexité stricte.

Proposition. *Si $p \geq 1$ et $q \geq 1$ alors la norme $\ell_{\mathbf{w};p,q}$ est convexe. La convexité stricte est obtenue pour $p > 1$ et $q > 1$.*

Les normes mixtes permettent de prendre en compte certaines structures qu'on peut trouver dans les signaux qu'on observe. Les propriétés classiques des normes, et en particulier la convexité, permettent de les utiliser efficacement dans un contexte de régression.

Afin d'illustrer le propos, considérons l'utilisation d'une norme $\ell_{\mathbf{w};21}$ sur le problème [4]. La norme $\ell_{\mathbf{w};21}$ définie sur une matrice $X \in \mathbb{R}^{I \times T}$, et dont les poids $\mathbf{w}$ ne dépendent que de l'indice i, est donnée par :

$$\|X\|_{\mathbf{w};21} = \sum_i \sqrt{\sum_t w_i |x_{i,t}|^2} \ .$$

Cela correspond à la somme des normes ℓ_2 de chaque ligne de X. La conséquence est qu'un X obtenu par minimisation de [4] est parcimonieux par ligne, i.e., les lignes de X contiennent soit uniquement des coefficients non-nuls soit uniquement des 0. Cette approche évite que les séries temporelles des sources soient irrégulières, ce qui

se produit lorsqu'on utilise une simple norme ℓ_1. Cette approche a été récemment proposée dans la littérature M/EEG dans (Ou *et al.*, 2009).

Historiquement et plus généralement, les normes mixtes sur deux niveaux ont été introduites dès les années 60 dans (Benedek *et al.*, 1961). Ces normes ont ensuite été étudiées plus particulièrement dans le cadre des espaces de modulation (Samarah *et al.*, 2006; Feichtinger, 2006). On retrouve aussi ce type de norme dans les espaces de Besov (Rychkov, 1999) ou de Triebel-Lizorkin (Grochenig *et al.*, 2000). On retrouve encore ces normes dans la communauté de l'apprentissage statistique. En particulier avec le Group-LASSO (Yuan *et al.*, 2006), qui correspond à la norme ℓ_{21}. On citera aussi l'utilisation de normes mixtes en classification (Szafranski *et al.*, 2010). Pour une revue plus complète sur les normes mixtes dans le cadre du traitement du signal on se reportera à (Kowalski, 2009) et (Kowalski *et al.*, 2009a), qui introduit notamment l'Elitist-LASSO avec la norme ℓ_{12}.

3.2. *Normes mixtes sur trois niveaux*

On s'intéresse particulièrement aux modèles où les sources peuvent être indicées par trois indices. Dans le cadre de la M/EEG, les trois indices peuvent correspondre à la localisation spatiale dans le cerveau, à l'indice temporel du signal, mais aussi à la condition expérimentale. Par exemple, pour les données somesthésiques, une condition correspond à un doigt de la main qui, lorsqu'il est stimulé, active une région précise du cortex sensoriel. Notons K le nombre de conditions. La matrice de mesures est obtenue par concatenation des mesures pour chaque condition. Elle est donnée par $M \in \mathbb{R}^{N \times KT}$. Les sources, dont les éléments sont indicés par (i,k,t), sont données par $X \in \mathbb{R}^{I \times KT}$; i indice l'espace, k la condition et t le temps.

On peut alors définir une norme mixte sur ces trois indices :

Définition (Normes mixtes sur trois niveaux)**.** *Soit* $\mathbf{x} \in \mathbb{R}^{IKT}$ *indicé par un triple indice* (i,k,t) *tel que* $\mathbf{x} = (x_{i,k,t})$. *Soit* $p,q,r \geq 1$ *et* $\mathbf{w} \in \mathbb{R}_{+,*}^{IKT}$ *une suite de poids strictement positifs, indicés par un triple indice* (i,k,t). *On appelle norme mixte de* $\mathbf{x}$ *la norme* $\ell_{\mathbf{w};pqr}$ *donnée par*

$$\|\mathbf{x}\|_{\mathbf{w};pqr} = \left(\sum_{i=1}^{I} \left(\sum_{k=1}^{K} \left(\sum_{t=1}^{T} w_{i,k,t} |x_{i,k,t}|^p \right)^{q/p} \right)^{r/q} \right)^{1/r} .$$

Le problème inverse considéré se présente de la manière suivante :

$$X^* = \arg\min_X \frac{1}{2}\|M - GX\|_F^2 + \frac{\lambda}{r}\|X\|_{\mathbf{w};pqr}^r \, , \lambda \in \mathbb{R}_+ \, . \qquad [9]$$

Pour l'application visée, on utilisera la norme mixte $\ell_{\mathbf{w};212}$. En choisissant de pénaliser les conditions expérimentales par une norme 1, tout en laissant une norme

2 sur les autres indices, on incite chaque source à être active pour peu de conditions. Avec l'exemple somesthésique choisi, une telle norme mixte favorise l'activation d'une seule région du cortex par condition expérimentale. Une telle norme permet par construction d'intégrer l'*a priori* que la stimulation de chaque doigt doit entrainer une activation neuronale à des endroits distincts. Ce point est illustré sur des simulations et des données MEG dans la section 5.

3.3. *Normes mixtes sur un nombre quelconque de niveaux*

L'extension de la définition d'une norme mixte sur un nombre quelconque de niveaux est immédiate, et l'on donne la définition dans un soucis d'exhaustivité :

Définition (Normes mixtes sur n niveaux)**.** *Soit $\mathbf{x} \in \mathbb{R}^N$ indicé par un nombre n d'indices $(i_1, \ldots, i_n)$ tel que $\mathbf{x} = (x_{i_1,\ldots,i_n})$. Soit $p_1, \ldots, p_n \geq 1$ et $\mathbf{w} \in \mathbb{R}^N_{+,*}$ une suite de poids strictement positifs, indicés par $(i_1, \ldots, i_n)$. On appelle norme mixte de $\mathbf{x}$ sur n niveaux la norme $\ell_{\mathbf{w};p_1,\ldots,p_n}$ donnée par*

$$\|\mathbf{x}\|_{\mathbf{w};p_1,\ldots,p_n} = \left(\sum_{i_n} \cdots \sum_{i_1} (w_{i_1,\ldots,i_n} |x_{i_1,\ldots,i_n}|^{p_1})^{p_2/p_1} \ldots \right)^{p_n/p_{n-1}} .$$

3.4. *Normes conjuguées*

On donne ici l'expression de la norme mixte conjuguée. Par la suite la notion de conjugué de Fenchel est utilisée afin de définir un critère d'arrêt des algorithmes itératifs utilisés. La définition du conjugué au sens de Fenchel est rappelée ci-après :

Définition (Conjugué de Fenchel)**.** *Soit f une fonction convexe. Le conjugué de fenchel de f, notée f^* est définie par :*

$$f^*(v) = \sup_u \langle v, u \rangle - f(u) .$$

v s'appelle la variable conjuguée.

Afin de calculer le conjugué de Fenchel d'une norme, on rappelle la notion de « norme duale ».

Définition (Norme duale)**.** *Soit $\|.\|$ une norme. On appelle norme duale la norme notée $\|.\|^*$ définie par*

$$\|u\|^* = \sup_v \{u^T v \mid \|v\| = 1\} .$$

Proposition (Norme duale d'une norme mixte)**.** *Le duale de la norme $\|.\|_{p_1,\ldots,p_n}$ est la norme $\|.\|_{p'_1 \ldots' p'_n}$ telle que , si $p_i > 1$, $\frac{1}{p_i} + \frac{1}{p'_i} = 1$, et si $p_i = 1$, $p'_i = \infty$.*

Démonstration. La démonstration est similaire à celle utilisée pour montrer que le dual d'une norme ℓ_p est la norme ℓ_q, telle que $\frac{1}{p} + \frac{1}{q} = 1$. Il suffit ici d'appliquer n fois l'inégalité de Hölder. ∎

Enfin, on donne l'expression de deux conjugués de Fenchel, l'une associée à une norme mixte, l'autre à une norme mixte élevée au carré.

Proposition (Conjugué de Fenchel d'une norme mixte)**.** *1) Le conjugué de Fenchel d'une norme est l'indicatrice de la norme conjuguée. Ainsi, le dual de la fonction* $u \mapsto \|u\|_{p_1,\dots,p_n}$ *est la fonction*

$$v \mapsto \chi_{\|v\|^*_{p_1,\dots,p_n}} = \begin{cases} 1 & si\ \|v\|_{p'_1,\dots,p'_n} \leq 1 \\ 0 & sinon\ . \end{cases}$$

où $\|v\|_{p'_1,\dots,p'_n}$ *est la norme mixte conjuguée.*

2) Le conjugué de Fenchel de la fonction $u \mapsto \frac{1}{2}\|u\|^2_{p_1,\dots,p_n}$ *est la fonction*

$$v \mapsto \frac{1}{2}\|v\|^2_{p'_1,\dots,p'_n}\ .$$

Pour une présentation plus détaillée du conjugué de Fenchel, et la démonstration des résultats ci-dessus, le lecteur peut se référer à (Boyd *et al.*, 2004) ou (Rockafellar, 1972)

4. Optimisation convexe

On rappelle dans cette section la notion d'opérateur de proximités, très employés dans les algorithmes utilisés pour minimiser une fonctionnelle non différentiable. On décrit deux de ces algorithmes, très utilisés en pratique. On discute enfin d'un critère d'arrêt pour ces algorithmes itératifs.

4.1. *Opérateurs de proximités*

La minimisation de la fonctionnelle [9] se fait par un algorithme itératif utilisant la notion d'opérateur de proximité (appelé aussi opérateur proximal) introduite par (Moreau, 1965). On s'intéressera à l'opérateur de proximité associé à la fonction de régularisation f_2. On rappelle tout d'abord la définition :

Définition (Opérateur de proximité)**.** *Soit $\phi : \mathbb{R}^P \to \mathbb{R}$ une fonction convexe semie-continue inférieurement. L'opérateur de proximité associé à ϕ et $\lambda \in \mathbb{R}_+$, noté $\text{prox}_{\lambda\phi} : \mathbb{R}^P \to \mathbb{R}^P$ est donné par*

$$\text{prox}_{\lambda\phi}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^P} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{x}) \ .$$

La proposition suivante rappelle les opérateurs de proximité des principales régularisations utilisées jusqu'alors :

Proposition (Quelques opérateurs de proximités utiles)**.** *Soit $\mathbf{x} \in \mathbb{R}^N$ et $\mathbf{y} \in \mathbb{R}^N$. Soit $\mathbf{w} \in {\mathbb{R}_+^*}^N$ un vecteur de poids.*
Energie minimale *On suppose que $\mathbf{x}$ n'est indicé que par un unique indice i. L'opérateur de proximité de la solution d'énergie minimale défini par :*

$$\text{prox}_{\lambda\|.\|_{\mathbf{w};2}}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \frac{1}{2}\lambda\|\mathbf{x}\|_{\mathbf{w};2}^2 \ ,$$

est donné, coordonnée par coordonnée, par :

$$x_i = \frac{y_i}{1+\lambda w_i} \ .$$

LASSO *On suppose que $\mathbf{x}$ n'est indicé que par un unique indice i. L'opérateur de proximité du LASSO défini par :*

$$\text{prox}_{\lambda\|.\|_{\mathbf{w};1}}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathbf{w};1} \ ,$$

est donné, coordonnée par coordonnée, par :

$$x_i = \frac{y_i}{|y_i|}\left(|y_i| - \lambda w_i\right)^+ \ .$$

Group-LASSO *On suppose que* $\mathbf{x}$ *est indicé par un double indice* (i,k)*. On suppose de plus que les poids ne changent qu'en fonction des groupes, c'est-à-dire que à* i *fixé,* $w_{i,k} = w_i$ *quelque soit* k*. L'opérateur de proximité du Group-LASSO défini par :*

$$\mathrm{prox}_{\lambda\|.\|_{\mathbf{w};21}}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathbf{w};21} \ ,$$

est donné, coordonnée par coordonnée, par :

$$x_{i,k} = y_{i,k}\left(1 - \frac{\lambda\sqrt{w_i}}{\|\mathbf{y}_i\|_2}\right)^+ .$$

Elitist-LASSO *On suppose que* $\mathbf{x}$ *est indicé par un double indice* (i,k)*. Soit l'opérateur de proximité de l'Elitist-LASSO défini par*

$$\mathrm{prox}_{\lambda\|.\|_{\mathbf{w};12}}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \frac{1}{2}\lambda\|\mathbf{x}\|_{\mathbf{w};12}^2 \ .$$

Soit, pour chaque i*,* $r_{i,k'_i} = y_{i,k'_i}/w_{i,k'_i}$ *ordonnés tels que, pour un* i *fixé,* $\forall k'_i, r_{i,k'_i+1} \leq r_{i,k'_i}$ *et les* $[y_{i,k}]$ *sont les* $y_{i,k}$ *ordonnés dans le même ordre que les* r_{i,k'_i}*. Soit* K_i *le nombre tel que*

$$\lambda \sum_{k'_i=1}^{K_i} w_{i,k'_i}^2 \left(r_{i,k'_i} - r_{i,K_i}\right) < r_{i,K_i} \leq \lambda \sum_{k'_i=1}^{K_i+1} w_{i,k'_i}^2 \left(r_{i,k'_i} - r_{i,K_i}\right)$$

L'opérateur de proximité est alors donné par

$$x_{i,k} = \frac{y_{i,k}}{|y_{i,k}|}\left(|y_{i,k}| - \frac{\lambda}{1+\lambda K_i}\sum_{k'_i=1}^{K_i}[y_{i,k'_i}]\right)^+ .$$

Ces opérateurs de proximités permettent de se faire une idée de la façon dont sont sélectionnés les coefficients de synthèse $\mathbf{x}$ en fonction des coefficients d'analyse $\mathbf{y}$. La solution d'énergie minimale est une simple pondération des coefficients d'analyse, tandis que le LASSO effectue un seuillage des coefficients, de façon indépendante. Le group-LASSO compare l'énergie des différents groupes à un seuil : seuls les groupes les plus énergétiques seront conservés, l'idée derrière le group-LASSO est «l'union fait la force». L'Elitist-LASSO à un comportement antagoniste : dans chaque groupe, seul les plus gros coefficients sont conservés, l'idée est de les «mettre en compétition».

Ces opérateurs de proximité permettent de déduire, après calculs, l'opérateur d'une norme mixte sur un nombre quelconque de niveaux, en se restreignant aux cas où les $p_i \in \{1,2\}$. L'application présentée utilisant plus spécifiquement la norme mixte sur trois niveaux $\ell_{\mathbf{w};212}$, l'opérateur proximal associé est explicité dans la proposition suivante :

Proposition (Opérateur de proximité de la norme mixte $\ell_{\mathbf{w};212}$)**.** *Soit* $\mathbf{y} \in \mathbb{R}^{IKT}$ *indicé par un triple indice* (i,k,t)*. Soit* $\mathbf{w}$ *une séquence de poids strictement positifs*

tels que $\forall t,\ w_{i,k,t} = w_{i,k}$. *Pour chaque* i, *soit* w_{i,k'_i}, $[y_{i,k'_i}] = \sqrt{w_{i,k'_i} \sum_t |y_{i,k'_i,t}|^2}$ *et* $r_{i,k'_i} = [y_{i,k'_i}]/w_{i,k'_i}$ *ordonnés tels que, pour un* i *fixé,* $\forall k'_i, r_{i,k'_i+1} \leq r_{i,k'_i}$. *Alors,* $\mathbf{x} = \mathrm{prox}_{\lambda\|\cdot\|^2_{\mathbf{w};212}}(\mathbf{y})$ *est donné coordonnée par coordonnée* (i,k,t) *par*

$$x_{i,k,t} = y_{i,k,t} \left(1 - \frac{\lambda\sqrt{w_{i,k}}}{1 + K_{\mathbf{w}_i}\lambda} \frac{\sum_{k'_i=1}^{K_i} [y_{i,k'_i}]}{\|\mathbf{y}_{i,k}\|_{\mathbf{w}_{i,k};2}} \right)^+ ,$$

avec $K_{\mathbf{w}_i} = \sum_{k_i=1}^{K_i} w^2_{i,k_i}$ *et l'indice* K_i *est le nombre tel que*

$$\lambda \sum_{k'_i=1}^{K_i} w^2_{i,k'_i} \left(r_{i,k'_i} - r_{i,K_i} \right) < r_{i,K_i} \leq \lambda \sum_{k'_i=1}^{K_i+1} w^2_{i,k'_i} \left(r_{i,k'_i} - r_{i,K_i} \right) \ ,$$

et $\mathbf{y}_{i,k}$ *la sous-suite issue de* $\mathbf{y}$ *pour* (i,k) *fixé.*

Démonstration. Une démonstration détaillée du calcul de l'opérateur de proximité associée à la norme ℓ_{212} est donnée à l'annexe A. ∎

4.2. *Algorithmes*

On utilise alors des algorithmes de type « seuillages itératifs », tels que ceux utilisés pour la pénalisation ℓ_1 par Daubechies *et al.* (Daubechies *et al.*, 2004), et que nous étendons aux problèmes pénalisés par normes mixtes (Kowalski, 2009). Ces algorithmes rentrent dans le cadre global des algorithmes proximaux étudiés récemment par Combettes et Wajs (Combettes *et al.*, 2005).

Ces algorithmes permettent de minimiser les fonctionnelles pouvant s'écrire sous la forme [2], sous les hypothèses suivantes :

– f_1 est une fonction propre, convexe semie-continue inférieurement et Lipschitz differentiable ;

– f_2 est une fonction propre, convexe semie-continue inférieurement (non nécessairement différentiable).

L'attache aux données [3] choisie est bien Lipschitz différentiable, avec pour gradient

$$\nabla f_1(X) = -G^*(M - GX) \ ,$$

et pour constante de Lipschitz $\|G^*G\|$.

L'algorithme le plus « simple » est l'algorithme 1 de seuillage itératif (ISTA pour Iterative Shrinkage/Thresholding Algorithm).

L'idée de la méthode est d'alterner la minimisation à la fois sur le terme d'attache aux données par des pas dans la direction opposée du gradient, et sur le terme de

Algorithm 1: ISTA

Initialisation : Choisir $X^{(0)} \in \mathbb{R}^{I \times KT}$ (par exemple $\mathbf{0}$).
repeat

$$X^{(k+1)} = \text{prox}_{\mu\lambda\ f_2}\left(X^{(k)} + \mu G^*(M - GX^{(k)})\right)$$

où $0 < \mu < 2\|G^*G\|^{-1}$.
until *convergence* ;

pénalisation par recours à l'opérateur proximal. Malheureusement cet algorithme peut s'avérer particulièrement lent à converger sur le problème considéré. La vitesse de convergence de l'algorithme est en $\mathcal{O}(1/k)$, où k est le nombre d'itérations.

On pourra lui préférer une version, moins intuitive, qui améliore fortement la vitesse de convergence par l'utilisation des schémas de Nesterov (Weiss, 2008) ou FISTA (Beck *et al.*, 2009) (Fast Iterative Shrinkage Thresholding Algorithm). La vitesse de convergence de ces algorithmes est en $\mathcal{O}(1/k^2)$, ce qui améliore significativement leur rapidité comparativement à ISTA. Le lecteur intéressé par les algorithmes rapides du premier ordre pourra se référer à (Tseng, 2009), qui présente une analyse unifiée et claire des avancées récentes dans ce domaine.

Algorithm 2: FISTA

Initialisation : $X^{(0)} \in \mathbb{R}^{I \times KT}$, $Z^{(1)} = X^{(0)}$, $\tau^{(1)} = 1$, $k = 1$, $\mu = 1/\|G^*G\|$
repeat

$$X^{(k)} = \text{prox}_{\mu\lambda f_2}\left(Z^{(k)} + \mu G^*(M - GZ^{(k)})\right)$$
$$\tau^{(k+1)} = \frac{1+\sqrt{1+4\tau^{(k)2}}}{2}$$
$$Z^{(k+1)} = X^{(k)} + \frac{\tau^{(k)}-1}{\tau^{(k)}}(X^{(k)} - X^{(k-1)})$$

until *convergence* ;

4.3. *Critère d'arrêt*

Les algorithmes utilisés sont itératifs et ne convergent pas en un nombre fini d'itérations vers la solution exacte. L'algorithme doit donc être stoppé lorsqu'un critère de convergence est vérifié. Le choix d'un tel critère détermine si la solution (nécessairement) approchée obtenue en sortie de l'algorithme est d'une qualité acceptable. Une telle décision est donc difficile et nécessite de construire un critère d'arrêt adapté. Idéalement, on aimerait minimiser la distance entre la solution estimée et le véritable minimiseur de la fonctionnelle qu'on cherche à optimiser.

Il existe divers critères de terminaison d'un algorithme itératif. On présente ici quelques critères classiques, plus ou moins judicieux. Un premier critère souvent rencontré est le suivant :

$$\|x^{(k)} - x^{(k-1)}\| < \varepsilon \,, \qquad [10]$$

où ε est un réel strictement positif. Ce critère se justifie par le fait que, si la suite des itérés produit par l'algorithme converge vers la solution, alors nécessairement la quantité définie dans [10] tend vers zéro. Cependant, cette quantité peut vite devenir très petite alors qu'on est encore très loin du minimiseur. On trouve de plus qu'en pratique, le choix d'un « bon » ε dépend fortement du paramètre λ de la fonctionnelle.

Une autre quantité utilisée en pratique modifie [10] en :

$$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k)}\|} < \varepsilon \,. \qquad [11]$$

Ce critère a les mêmes défauts que le précédent et reflète donc mal la convergence des itérés.

Lorsque la fonctionnelle $\mathcal{F}$ est différentiable un critère naturel est de regarder si le gradient s'annule. Cela se traduit par

$$\|\nabla \mathcal{F}(X^{(k)})\| < \varepsilon \,. \qquad [12]$$

Malheureusement les fonctionnelles faisant intervenir une norme ℓ_1 ne sont pas différentiables et un tel critère est donc exclu.

Un bien meilleur critère lorsqu'on s'intéresse à un problème d'optimisation convexe, tels que ceux mentionnés dans ce travail, est basé sur un résultat fondamental qui garantit qu'à l'optimum le « saut de dualité » du problème est nul. Ce résultat basé sur la dualité Lagrangienne ou de Fenchel passe par l'introduction du problème, dit « dual », associé à notre problème d'optimisation qui lui est dénommé problème « primal ». Le saut de dualité est défini comme l'écart entre le minimum de la fonctionnelle primale $\mathcal{F}_p$ et le maximum du problème dual qui lui est concave. En construisant explicitement une fonctionnelle duale $\mathcal{F}_d$ et en associant à la variable primale $X^{(k)}$ une variable duale notée $Y^{(k)}$, on peut calculer un saut de dualité $\eta^{(k)}$ défini alors comme :

$$\eta^{(k)} = \mathcal{F}_p(X^{(k)}) - \mathcal{F}_d(Y^{(k)}) \; \geq 0 \,.$$

A l'optimum $\eta^{(k)}$ est nul. Ainsi en retournant un X et un η l'algorithme garantit que $\mathcal{F}_p(X)$ est au plus à une distance $\mathcal{F}_p(X) + \eta$ de l'optimum. Un bon critère d'arrêt peut alors être :

$$\mathcal{F}_p(X^{(k)}) - \mathcal{F}_d(Y^{(k)}) < \varepsilon \,. \qquad [13]$$

Reste le problème de trouver une expression pour $\mathcal{F}_d$ et de savoir associer une variable duale Y à une variable primale X pour les problèmes de la forme [4] qui sont considérés dans ce travail. On utilise ici la notion de dualité de Fenchel, et en particulier le théorème suivant (Rockafellar, 1972) :

Théorème. *Soit f une fonction convexe de $\mathbb{R}^m$ et g une fonction concave de $\mathbb{R}^n$. Soit G un opérateur linéaire de $\mathbb{R}^m$ dans $\mathbb{R}^n$, alors*

$$\inf_{x\in\mathbb{R}^m}\{f(x)-g(Gx)\} = \sup_{u\in\mathbb{R}^n}\{g^*(u)-f^*(G^*u)\}\ ,$$

où f^* (resp. g^*) est le conjugué de Fenchel de f (resp. g), et G^* l'opérateur adjoint de G.

De plus, les conditions de Karush-Khun-Tucker (KKT) sont équivalentes à

$$\begin{aligned} f(x)+f^*(G^*u) &= \langle x, G^*u\rangle\ ,\\ g(Gx)+g^*(u) &= \langle Gx, u\rangle\ . \end{aligned}$$

Ainsi, si on considère le problème général [4], le problème dual associé est :

$$\max_Y -\frac{1}{2}\|Y\|_2^2 + \mathrm{Tr}(Y^TM) - f_2^*(G^*Y)\ , \qquad [14]$$

où f_2^* est le conjugué de Fenchel de λf_2 et Tr correspond à la trace d'une matrice. Or les conjugués de Fenchel des normes mixtes et des normes mixtes élevées au carré ont été calculées à la section 3 et correspondent respectivement à l'indicatrice de la norme conjuguée et à la norme conjuguée élevée au carré. Le calcul des conjugués, que l'on ne détaille pas ici, est obtenu essentiellement grâce à la proposition 3.4 de la section 3. Le point $Y^{(k)}$ du problème dual est obtenu par construction à partir de $X^{(k)}$ à partir des conditions KKT du théorème 4.3, ce qui donne ici

$$Y^{(k)} = M - GX^{(k)}\ ,$$

que l'on projète si nécessaire pour satisfaire la contrainte f_2^*, dans le cas où f_2^* correspond à une indicatrice. On peut alors utiliser le critère [13] comme critère d'arrêt de nos algorithmes.

5. Résultats

Cette section présentent deux résultats obtenus en localisation de sources M/EEG. Le premier est un résultat issu d'une simulation, afin de pouvoir mesurer de manière objective les performances de différentes approches. Le second est une application à un problème réel.

5.1. *Simulation*

De légères stimulations électriques des doigts de la main génèrent une activité neuronale mesurable par M/EEG. À chaque doigt de la main correspond une zone fonctionnelle différente, zones qui sont de plus limitrophes au sein du cortex sensoriel primaire pour deux doigts contigus. Ces connaissances neuroanatomiques montrent que le cortex sensoriel primaire possède une structure que nous proposons d'utiliser comme *a priori* dans le problème inverse à l'aide d'une norme mixte $\ell_{\mathbf{w};212}$, qui pénalise le chevauchement entre deux régions corticales par une norme ℓ_1.

Pour illustrer cela, on reproduit une partie de l'organisation du cortex sensoriel primaire (S1) (Penfield *et al.*, 1950). On simule une activité sur trois régions corticales distinctes contiguës (voir la figure 2a), qui peuvent correspondre à la localisation des zones sensorielles primaires de trois doigts de la main droite. On a ensuite généré des mesures synthétiques bruités par un bruit blanc gaussien additif. L'amplitude d'activation de la région la plus temporale (en rouge sur la figure 2), qui peut correspondre au pouce, est deux fois plus grande que les deux autres régions. En pratique, l'amplitude des sources est différente selon les conditions. Les résultats obtenus par l'utilisation de la norme mixte $\ell_{\mathbf{w};212}$ [9], d'une norme $\ell_{\mathbf{w};F}$ [7] et d'une norme $\ell_{\mathbf{w};1}$ sont ensuite comparés.

On évalue chaque méthode pour différents niveaux de bruit (donnés par le rapport signal à bruit (SNR)). Une fois une solution au problème inverse calculée, le maximum d'amplitude d'une source à travers les conditions permet d'attribuer une source à une condition. La performance est mesurée en comptant le pourcentage de sources mal étiquetées (c'est-à-dire, attribuées à la mauvaise condition). La figure 1 montre le pourcentage d'erreur obtenu pour les trois pénalisations. Chaque résultat est présenté avec une barre d'erreur obtenue par calcul de la variance du résultat sur 10 réalisations de l'expérience. On peut voir que la norme mixte $\ell_{\mathbf{w};212}$ donne toujours les meilleurs résultats. La performance de la norme $\ell_{\mathbf{w};1}$ chute rapidement lorsque le SNR diminue, ce qui est cohérent avec les précédentes observations faites par la communauté M/EEG.

Afin d'avoir une comparaison la plus juste possible, le paramètre λ a été ajusté dans chaque cas, de manière à ce que la norme du résidu $\|M - GX^*\|_F$ soit égale à la norme du bruit ajouté (qui est connu dans les simulations).

Les résultats sont aussi illustrés sur les figures 2b et 2c sur une région d'intérêt (RI) autour du cortex somatosensoriel primaire gauche. On peut voir que l'étendue de

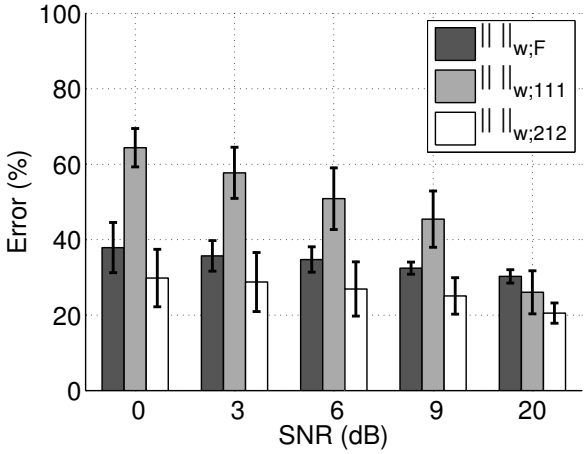


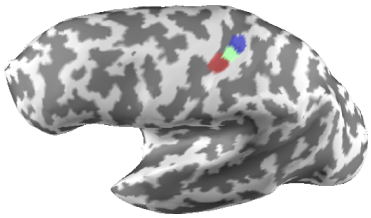
FIGURE 1. *Évaluation des estimées par* $\|.\|_{\mathbf{w};F}$ *vs.* $\|.\|_{\mathbf{w};212}$ *vs.* $\|.\|_{\mathbf{w};111}$ *sur des données somesthésiques synthétiques. L'erreur représente le pourcentage des sources mal identifiés.*

la région la plus temporale obtenue avec la pénalité $\|.\|_{\mathbf{w};F}$, est largement surestimée, contrairement aux résultats obtenus avec la pénalité $\|.\|_{\mathbf{w};212}$. Bien qu'imparfaits, les résultats obtenus à l'aide de la norme $\|.\|_{\mathbf{w};212}$ restitue la bonne organisation du cortex sensoriel primaire en fournissant des zones d'activations dont l'étendue spatiale est correcte. Cette observation est d'ailleurs confirmée par les résultats de simulations présentés sur la figure 2 pour différentes valeurs de SNR.

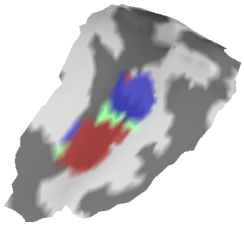
Les deux algorithmes ISTA (Algorithme 1) et FISTA (Algorithme 2) ont été utilisés. En général, afin d'illustrer la vitesse de convergence, l'évolution de la valeur de la fonctionnelle est donnée en fonction du nombre d'itération, car les résultats théoriques portent sur cette quantité. Cependant, comme en pratique nous utilisons le saut de dualité comme critère de convergence, nous avons choisis de montrer sur la figure 3 la vitesse de convergence en terme de saut de dualité de ces deux algorithmes, en fonction du nombre d'itération.

5.2. *Données réelles*

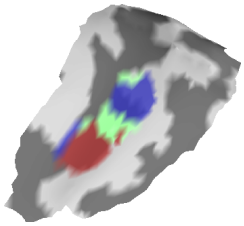
Les données ont été enregistrées avec une fréquence de 1250 Hz sur une machine MEG ayant 151 capteurs. Les doigts ont été stimulés électriquement de façon aléatoire. Les données ont été obtenues par moyennage de 400 répétitions. Le problème direct a été calculé avec un modèle sphérique et un maillage cortical contenant environ 55000 sources afin de permettre une bonne précision spatiale. Le problème inverse a été calculé sur une fenêtre de temps de 5 ms environ 45 ms après stimulation, ce



(a) Données simulées



(b) Résultat avec $\|.\|_{\mathbf{w};F}$ (RI)



(c) Résultat avec $\|.\|_{\mathbf{w};212}$ (RI)

FIGURE 2. *Illustration des résultats sur le cortex sensoriel primaire (S1) (SNR = 20dB). Les régions actives voisines reproduisent l'organisation de S1.*

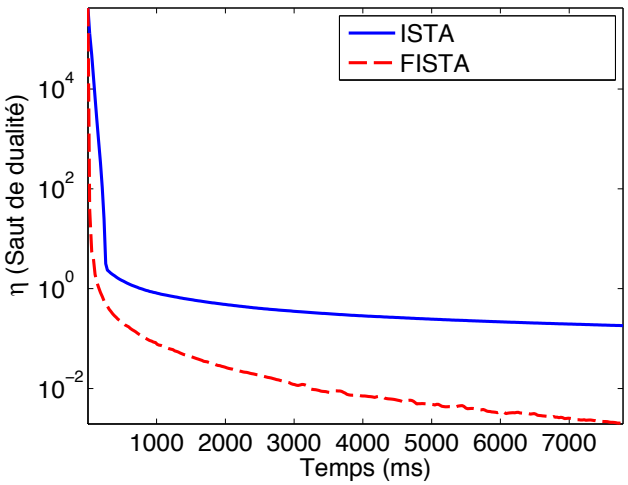
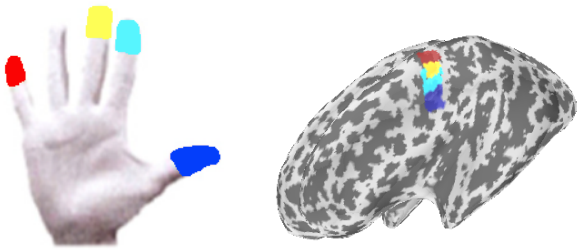


FIGURE 3. *Evolution comparée de la valeur du saut de dualité pour ISTA et FISTA.*

qui correspond à la période d'activité maximale. Les colonnes $G_{\cdot i}$ de G n'étant pas normalisées, les poids sont donnés par $w_{i,k} = \|G_{\cdot i}\|_2$.

Une fois une solution au problème inverse calculée, la labélisation des zones actives a été faite en attribuant à chaque source le label correspondant à la condition pour laquelle l'activité reconstruite était la plus grande. Sont représentés sur la figure 4 les résultats obtenus à la fois pour la norme classique $\|.\|_{\mathbf{w},F}$ et pour $\|.\|_{\mathbf{w},212}$. Afin de rendre la comparaison équitable le paramètre λ a été choisi pour chaque norme afin d'obtenir à l'optimum un même terme d'attache aux données ayant été estimé sur une période de repos précédant la stimulation. Pour chaque condition, la région active a été délimitée par calcul de la plus grande composante connexe. On constate qu'en injectant l'*a priori* structuré à l'aide de la norme $\ell_{\mathbf{w};212}$ l'organisation du cortex sensoriel primaire est correctement retrouvée alors que la norme classiquement utilisée en M/EEG échoue en surestimant l'étendue de la zone active pour le deuxième doigt (l'index).



(a) Doigts labélisés (b) Résultat avec $\|.\|_{\mathbf{w},212}$

(c) Résultat avec $\|.\|_{\mathbf{w},212}$ (zoom) (d) Résultat avec $\|.\|_{\mathbf{w},F}$ (zoom)

FIGURE 4. *Résultats de labélisation du cortex sensoriel primaire par l'utilisation d'une norme mixte $\ell_{\mathbf{w};212}$ sur des données de MEG.*

6. Conclusion

Nous avons montré comment l'utilisation d'une norme mixte adéquate permet de structurer l'*a priori* intervenant dans la régularisation d'un problème inverse. Le cadre choisi, c'est à dire l'optimisation convexe, permet d'utiliser des algorithmes du premier ordre qui sont simples et rapides à mettre en oeuvre tout en fournissant une vitesse de convergence pratique en rapport avec les vitesses de convergence théoriques. Nous avons détaillé les opérateurs de proximités utilisés pour implémenter les algorithmes. Ces opérateurs correspondent en pratique à des opérations de seuillages. De plus, les algorithmes itératifs présentés sont fournis avec un critère d'arrêt donnant des garanties sur l'optimalité des solutions obtenues.

L'avantage des normes mixtes est montré à la fois sur un exemple synthétique et sur des données réelles MEG. Les résultats obtenus ont permis ainsi de cartographier le cortex sensoriel primaire d'un individu avec plus de précision que des méthodes basées sur des a priori non structurés classiquement utilisés dans le cadre de l'étude de données M/EEG. La cartographie du cortex sensoriel primaire peut notamment être utilisée dans le cadre d'études cliniques comme cela a déjà été fait pour l'étude de patients souffrant de dystonie (Meunier *et al.*, 2001). Plus généralement, les méthodes présentées peuvent servir à la cartographie fonctionnelle, notamment celle des aires visuelles comme le montre des résultats préliminaires présentés dans (Gramfort, 2009). Enfin, nous avons choisi de présenter les normes mixtes à travers une application spécifique : la localisation de sources en M/EEG, mais il existe de nombreux autres problèmes pour lesquels l'emploi de telles normes est justifié, notamment en séparation de sources.

Signalons qu'il existe une infinité de façon de régulariser un problème inverse. Rien qu'en se restreignant au cadre de l'optimisation convexe, on peut utiliser des termes de régularisation « composites », qui font intervenir plusieurs *a priori* indépendants, du type :

$$f_2(X) = \mu g_1(X) + (1-\mu)\mu g_2(X)\ , \mu \in [0,1]\ .$$

Par exemple, l'elastic-net utilise une régularisation faisant intervenir une norme ℓ_1 et une norme ℓ_2 (Zou *et al.*, 2005) :

$$f_2(X) = \mu \|X\|_1 + (1-\mu)\|X\|_2^2\ ,$$

ce qui a pour effet pratique de conserver des petits groupes de coefficients corrélés, et également de garantir l'unicité de la solution en rendant la fonctionnelle considérée strictement convexe.

Les algorithmes présentés ici, et beaucoup d'autres, ont été développés sous MATLAB, dans le projet EMBAL (Electro-Magnetic Brain Activity Localization)[1]. Ce projet à pour but de fournir toute une série de méthodes pour la localisation de sources en

1. `http://embal.gforge.inria.fr/`

M/EEG, mais propose aussi la plupart des algorithmes d'optimisation convexe servant à minimiser une fonction du type [2]. Une large gamme d'opérateurs de proximité y sont aussi implémentés.

A. Calcul de l'opérateur de proximité de la norme ℓ_{212}

On cherche :

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathbf{w};212}^2 \ ,$$

avec

$$\|\mathbf{x}\|_{\mathbf{w};212}^2 = \sum_i \left(\sum_j \left(\sum_k w_{i,k}|x_{i,k,t}|^2 \right)^{1/2} \right)^2 \ .$$

On dérive par rapport à $x_{i,k,t}$ et on obtient le système d'équations variationnelles :

$$\begin{aligned} |x_{i,k,t}| &= |y_{i,k,t}| - \lambda w_{i,k}|x_{i,k,t}|(\sqrt{w_{i,k}}\|\mathbf{x}_{i,k}\|_2)^{-1}\|\mathbf{x}_i\|_{\mathbf{w};21} \\ \arg(x_{i,k,t}) &= \arg(y_{i,k,t}) \ . \end{aligned}$$

ce qui donne :

$$\begin{aligned} |x_{i,k,t}| \left(1+\lambda\sqrt{w_{i,k}}\|\mathbf{x}_{i,k}\|_2^{-1}\|\mathbf{x}_i\|_{\mathbf{w};21}\right) &= |y_{i,k,t}| \qquad [15] \\ |x_{i,k,t}|^2 \left(1+\lambda\sqrt{w_{i,k}}\|\mathbf{x}_{i,k}\|_2^{-1}\|\mathbf{x}_i\|_{\mathbf{w};21}\right)^2 &= |y_{i,k,t}|^2 \ . \end{aligned}$$

On somme sur t, et on obtient :

$$\begin{aligned} \|\mathbf{x}_{i,k}\|_{\mathbf{w}_{i,k};2} \left(1+\lambda\sqrt{w_{i,k}}\|\mathbf{x}_{i,k}\|_{\mathbf{w}_{i,k};2}^{-1}\|\mathbf{x}_i\|_{\mathbf{w};21}\right) &= \|\mathbf{y}_{i,k}\|_2 \\ \|\mathbf{x}_{i,k}\|_2 + \lambda\sqrt{w_{i,k}}\|\mathbf{x}_i\|_{\mathbf{w};2,1} &= \|\mathbf{y}_{i,k}\|_2 \ . \qquad [16] \end{aligned}$$

La solution de cette dernière équation est donnée par la proposition 4.1 (c'est exactement l'équation variationnelle obtenue pour l'Elitist-Lasso) :

$$\|\mathbf{x}_{i,k}\|_2 = \left(\|\mathbf{y}_{i,k}\|_2 - \frac{\lambda\sqrt{w_{i,k}}}{1+K_{\mathbf{w}_i}} \sum_{j=1}^{K_i} [y_{i,k_i'}] \right)^+ \ . \qquad [17]$$

où $K_{\mathbf{w}_i} = \sum_{i=1}^{K_i} w_{i,k}$.

On revient à [15] qu'on réécrit :

$$\begin{aligned} |x_{i,k,t}| &= \frac{|y_{i,k,t}|}{\left(1+\lambda\sqrt{w_{i,k}}\|\mathbf{x}_{i,k}\|_2^{-1}\|\mathbf{x}_i\|_{\mathbf{w}_i;21}\right)} \\ &= \frac{|y_{i,k,t}|\|\mathbf{x}_{i,k}\|_2}{\left(\|\mathbf{x}_{i,k}\|_2+\lambda\sqrt{w_{i,k}}\|\mathbf{x}_i\|_{\mathbf{w}_i;21}\right)} \ . \end{aligned}$$

En utilisant [16], on obtient simplement :

$$|x_{i,k,t}| = \frac{|y_{i,k,t}|\|\mathbf{x}_{i,k}\|_2}{\|\mathbf{y}_{i,k}\|_2} \ .$$

Et donc, en injectant [17] dans cette dernière équation on a

$$
\begin{aligned}
|x^*_{i,k,t}| &= \frac{|y_{i,k,t}| \left(\|\mathbf{y}_{i,k}\|_2 - \frac{\lambda\sqrt{w_{i,k}}}{1+K_{\mathbf{w}_i}} \sum_{j=1}^{K_i} [y_{i,k'_i}] \right)^+}{\|\mathbf{y}_{i,k}\|_2} \\
&= |y_{i,k,t}| \left(1 - \frac{\lambda\sqrt{w_{i,k}}}{1+K_{\mathbf{w}_i}} \frac{\sum_{j=1}^{K_i} [y_{i,k'_i}]}{\|\mathbf{y}_{i,k}\|_2} \right)^+ .
\end{aligned}
$$

Remerciements

Les auteurs remercient le Dr. Sabine Meunier de l'hôpital La Salpêtrière, INSERM, Paris, pour sa permission d'utiliser une partie de ses données MEG.

Ce travail a été financé en parti par la Fondation EADS - Contrat Odyssee-EADS 2118 et le contrat ANR ViMAGINE ANR-08-BLAN-0250-02.

B. Bibliographie

Baillet S., Mosher J., Leahy R., « Electromagnetic brain mapping », *IEEE Signal Processing Magazine*, vol. 18(6), p. 14-30, 2001.

Beck A., Teboulle M., « A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems », *SIAM Journal on Imaging Sciences*, vol. 2, n° 1, p. 183-202, 2009.

Benedek A., Panzone R., « The space L^p with mixed norm », *Duke Mathematical Journal*, vol. 28, p. 301-324, 1961.

Bobin J., Starck J.-L., Moudden Y., Fadili M., « Blind Source Separation : The Sparsity Revolution », *in* P. Hawkes (ed.), *Advances in Imaging and Electron Physics*, Academic Press, Elsevier, p. 221-298, 2008.

Boyd S., Vandenberghe L., *Convex Optimization*, Cambridge University Press, March, 2004.

Candès E. J., Tao T., « The Dantzig selector : statistical estimation when p is much larger than n », *Annals of Statistics*, vol. 35, p. 2313-2351, 2005.

Combettes P. L., Wajs V. R., « Signal recovery by proximal forward-backward splitting », *Multiscale Modeling and Simulation*, vol. 4, n° 4, p. 1168-1200, November, 2005.

Dale A., Liu A., Fischl B., Buckner R., « Dynamic statistical parametric neurotechnique mapping : combining fMRI and MEG for high-resolution imaging of cortical activity », *Neuron*, vol. 26, p. 55-67, 2000.

Dale A., Sereno M., « Improved Localization of Cortical Activity By Combining EEG and MEG with MRI Cortical Surface Reconstruction », *Journal of Cognitive Neuroscience*, vol. 5, p. 162-176, Jan, 1993.

Daubechies I., Defrise M., De Mol C., « An iterative thresholding algorithm for linear inverse problems with a sparsity constraint », *Commun. Pure Appl. Math.*, vol. 57, n° 11, p. 1413 - 1457, Aug, 2004.

Daudet L., Molla S., Torrésani B., « Towards a hybrid audio coder », *in* J. P. Li (ed.), *International Conference Wavelet analysis and Applications*, Chongqing, China, p. 13-24, 2004.

Donoho D. L., « Compressed Sensing », *IEEE Transactions on Information Theory*, vol. 52, n° 4, p. 1289-1306, april, 2006.

Dupé F.-X., Fadili M., Starck J.-L., « A proximal iteration for deconvolving Poisson noisy images using sparse representations », *IEEE Transactions on Image Processing*, vol. 18, n° 2, p. 310-321, 2009.

Feichtinger H. G., « Modulation Spaces : Looking Back and Ahead », *Sampling Theory in Signal and Image Processing*, vol. 5, n° 3, p. 109-140, 2006.

Févotte C., Torrésani B., Daudet L., Godsill S. J., « Sparse linear regression with structured priors and application to denoising of musical audio », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, n° 1, p. 174-185, 2008.

Gramfort A., Mapping, timing and tracking cortical activations with MEG and EEG : Methods and application to human vision, PhD thesis, 2009.

Grochenig K., Samarah S., « Non-linear approximation with local Fourier bases », *Constr. Approx.*, vol. 16, p. 317-332, 2000.

Kowalski M., « Sparse Regression Using Mixed Norms », *Appl. Comput. Harmon. Anal.*, vol. 27, n° 3, p. 303-324, 2009.

Kowalski M., Torrésani B., « Random Models for Sparse Signals Expansion on Unions of Basis with Application to Audio Signals », *IEEE Transactions on Signal Processing*, 2008.

Kowalski M., Torrésani B., « Sparsity and persistence : mixed norms provide simple signals models with dependent coefficients », *Sig Imag Video Process*, vol. 3, n° 3, p. 251-264, 2009a.

Kowalski M., Vincent E., Gribonval R., Beyond the Narrowband Approximation : Wideband Convex Methods for Under-Determined Reverberant Audio Source Separation, Technical report, 2009b.

Meunier S., Garnero L., Ducorps A., Mazières L., Lehéricy S., du Montcel S., Renault B., Vidailhet M., « Human brain mapping in dystonia reveals both endophenotypic traits and adaptative reorganization », *Annals of Neurology*, vol. 50, p. 521-527, 2001.

Moreau J.-J., « Proximité et dualité dans un espace hilbertien », *Bull. Soc. Math. France*, vol. 93, p. 273-299, 1965.

Mosher J., Lewis P., Leahy R., « Multiple Dipole Modeling and Localization from Spatio-Temporal MEG data », *IEEE Transactions on Biomedical Engineering*, vol. 39, n° 6, p. 541-553, 1992.

Ou W., Hämaläinen M., Golland P., « A Distributed Spatio-Temporal EEG/MEG Inverse Solver », *Neuroimage*, vol. 44, p. 932-946, 2009.

Pascual-Marqui R. D., Michel C. M., Lehman D., « Low resolution electromagnetic tomography : A new method for localizing electrical activity of the brain », *Psychophysiology*, vol. 18, p. 49-65, 1994.

Penfield W., Rasmussen T., *The Cerebral Cortex of Man : A Clinical Study of Localization of Function*, Macmillan, 1950.

Rockafellar R., *Convex Analysis*, Princeton University Press, 1972.

Rychkov V. S., « On restrictions and extensions of the Besov and Triebel–Lizorkin spaces with respect to Lipschitz domains », *Journal of London Mathematical Society*, vol. 60, n° 1, p. 237-257, 1999.

Samarah S., Salman R., « Local Fourier bases and modulation spaces », *Turkish Journal of Mathematics*, vol. 30, n° 4, p. 447-462, 2006.

Szafranski M., Grandvalet Y., Rakotomamonjy A., « Composite Kernel Learning », *Machine Learning*, vol. , p. 1-33, 2010.

Tseng P., Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization, Technical report, 2009.

Vincent E., Gribonval R., Pumbley M., « Oracle estimators for the benchmarking of source separation algorithms », *Sig. Process.*, vol. 87, n° 8, p. 1933 - 1950, Aug, 2007.

Weiss P., Algorithmes rapides d'optimisation convexe. Applications à la reconstruction d'images et à la détection de changements., PhD thesis, Université de Nice Sophia-Antipolis, Novembre, 2008.

Wipf D., Nagarajan S., « A unified bayesian framework for MEG/EEG source imaging », *Neuroimage*, vol. 44, n° 3, p. 947 - 966, Feb, 2009.

Yuan M., Lin Y., « Model selection and estimation in regression with grouped variables », *Journal of the Royal Statistical Society Serie B*, vol. 68, n° 1, p. 49-67, 2006.

Zou H., Hastie T., « Regularization and variable selection via the elastic net », *Journal of the Royal Statistical Society Series B*, Jan, 2005.